표본추출법

- 표본추출이론과 추정

변 종 석
(한신대학교 응용통계학과)

3. 표본추출방법

- 표본추출방법 결정 시 고려 사항
  - 모집단
  - 추출단위
  - 표본추출과정
  - 표본크기
  - 추정 및 분석방법

W.E. Deming
" 표본추출이란 전체를 대신하는 일부를 선택하는 것이라기 보다는 확률적 근거에 의해 유용한 통계적 정보의 신뢰성을 측정하고 조절하는 과학이며 예술이다 "

- 표본추출법
  - 비확률추출법(non-probability sampling)
  - 확률추출법(probability sampling)

1. 비확률추출법

- 비확률추출법의 특성
  - 조사자의 주관이나 경험에 의해 편의적으로 표본을 추출하는 방법
  - 표본추출이 간편하고 경제적임
  - 조사자의 주관적 개입으로 표본을 추출하므로 결과의 일반화에 어려움이 존재
  - 추정 및 분석 과정의 정확성을 평가할 수 없음이 단점(과학적 조사 방법으로서의 한계 존재)
    - 모든 통계분석방법의 기본적인 자료는 확률추출에 의한 표본으로부터 수집한 자료를 가정함

- 비확률추출법의 종류
  - 간편추출법 (convenience sampling)
  - 판단추출법 (judgement sampling)
  - 할당추출법 (quota sampling)
  - 눈덩이추출법 (snowball sampling)

1) 간편추출법 (convenience sampling)

- 정의 : 조사원의 자의적인 판단에 위해 간편한 방법으로 표본을 추출하는 방법
- 예 :
  - 거리에서 지나는 사람을 대상으로 임의 선택하여 면접조사
  - 표본추출과정없이 ARS를 이용한 전화여론조사
  - 표본추출과정없이 인터넷을 이용한 여론조사 및 소비자조사

2) 판단추출법(judgement sampling ; purposive sampling)

- 정의 : 연구자 혹은 조사원의 전문적인 판단이나 지식, 경험에 의해 의도적으로 표본을 선택하는 방법
- 특성
  - 표본크기가 작거나 정성조사에서 유용한 방법
  - 추정량의 정확성을 평가하지 못함
- 예 :
  - Opinon leader group 조사
  - 전체를 대표한다고 생각되는 일부 가구만을 선정하는 조사
  - 평균에 해당하는 일부 대상만을 선정하는 조사

3) 할당추출법 (quota sampling)

- 정의 : 알려진 모집단의 구조와 동일한 구성 비율을 유지하도록 표본을 선정하는 방법
- 특성
  - 적은 비용, 표본 확보의 편리성때문에 단기간 조사에 적합한 방법
  - 조사 목적과 관련이 높은 중요 변수만을 고려하여 표본을 선택하는 방법
  - 조사 대상 특성 중 큰 오차를 유발하는 변수에 대해 오차를 감소하도록 표본 선정이 가능한 방법
  - 비전문가에게 표본 구조에 대한 이해가 용이한 추출 방법(주의 : 비례배분의 확률추출법과 구조가 동일)
  - 표본추출 과정에 비확률적 요소가 포함된 추출 방법
- 예 : 전화조사 및 여론조사에서 널리 채택하는 있는 조사

4) 눈덩이추출법(snowball sampling)

- 특정 표본으로부터 해당 집단에 속한 다른 대상을 소개받아서 표본으로 선정하는 방법
- 특성
  - 조사 대상이 희귀하거나 접근이 어려운 경우에 유용한 방법
  - 추출틀 작성이 어려운 경우에 유용한 추출 방법
  - 비확률적인 적응탐색적 추출 방법 (an adaptive searching sampling)
- 예 :
  - 희귀 유전병 조사 등

2. 확률추출법

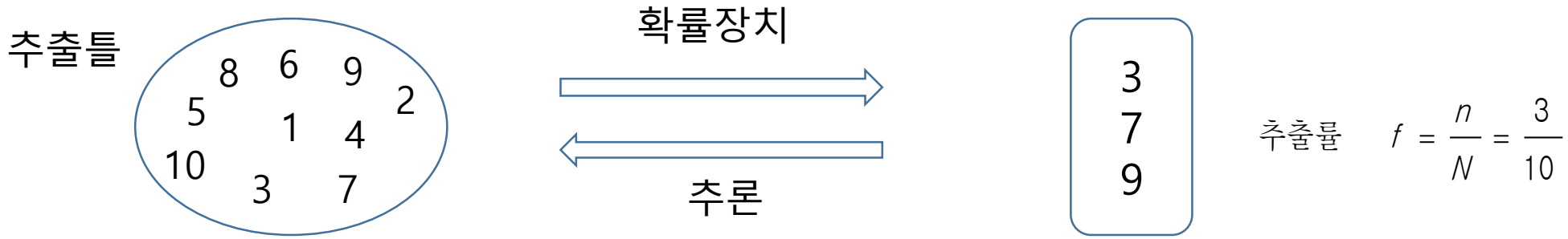
- 확률추출법의 특성
  - 모집단에 속한 모든 추출단위에 대해서 사전에 정해진 추출확률에 의해 표본을 추출하는 방법
  - 조사 결과로부터 모집단 전체에 대한 일반화가 가능한 추출 방법
  - 추정량의 정확성 및 신뢰성에 대한 평가가 가능한 방법
    - 모든 통계분석방법을 이용한 분석이 가능하고, 그 결과에 대해 일반화가 가능함
  - 추정 및 분석 과정이 다소 복잡

- 전통적인 확률추출법의 종류
  - 단순확률추출법 (SRS : simple random sampling)
  - 층화추출법 (STS : stratified sampling)
  - 계통추출법 (SYS : systematic sampling)
  - 집락추출법 (CLS : cluster sampling)

1) 단순확률추출법 (SRS : simple random sampling)

- 정의 : 모집단 대상에게 일련번호를 부여하여 적절한 확률장치를 통해 무작위로 표본을 추출하는 방법
- 추출방법
  - 모집단 대상(추출틀의 추출단위)에게 일련번호를 부여하고
  - 확률장치(난수표, 난수 생성 프로그램)를 통해 n개의 번호를 무작위로 추출하여
  - 확률장치를 통해 추출된 번호와 일치하는 대상을 표본으로 선정
- 특성
  - 모집단을 대표하는 가장 이상적인 방법으로 추출틀만 확보되면 적용이 가능한 방법
  - 추출단위의 추출확률이 동일하다고 가정 : 모든 대상이 표본으로 추출될 기회가 동일(=가중치 동일)
  - 확률추출법의 기본 방법으로 다른 확률추출법의 기초 추출방법으로 활용
  - 추출 방법 : 복원추출(WR: With Replacement)과 비복원추출 (WOR: WithOut Replacement)
  - 단점
    - N이 매우 크면 적용하기 어려운 방법 : 대규모조사에 적용하기 어려움
    - 추출단위들이 이질적인 경우 편향된 표집 구성의 가능성이 존재
- 예 :
  - RDD

- 표본자료(정보)에 대한 이해



- 추출 : N=10개에서 확률추출과정으로 n=3개 추출
- 추론 : 표본 n=3개 자료로부터 모집단 N=10개의 정보를 파악하게 됨
- 표본단위의 정보 : 표본 추출율(sampling fraction)이 3/10이므로 표본 1개는 모집단 기준 10/3의 정보를 가지는 것으로 설명하면 3개 표본자료로부터 모집단 10개 정보 파악이 가능

- 가중치(weight)
  - 가중치란 ? 표본단위가 가지고 있는 모집단 정보의 양을 표현해 주는 모집단으로의 확장계수(expansion coefficient)를 의미
  - 가중치=추출율의 역수=1/fraction
  - 가중치의 역할 : 표본자료를 모집단 관점에서 추정 혹은 분석하도록 해 주는 역할

$$w_i = \frac{1}{f} \xrightarrow{SRS} w_i = \frac{N}{n} \Rightarrow \sum_{i=1}^{n} w_i = N$$

- **추정식 : 모평균**

  - 모평균 추정 식: 표본평균

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} \left(\frac{N}{n}\right) y_i}{n\left(\frac{N}{n}\right)} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

  - 표본평균의 분산 추정량(WOR 가정)

$$\hat{V}(\bar{y}) = \frac{N-n}{N}\frac{s^2}{n}$$

  - 추정의 정도

    . 추정오차한계(=신뢰구간의 폭)

$$\hat{B} = z\sqrt{\hat{V}(\bar{y})} \approx 2\sqrt{\frac{N-n}{N}\frac{s^2}{n}}$$

    . 상대표준오차

$$\hat{RSE} = \frac{\sqrt{\hat{V}(\bar{y})}}{\bar{y}} = \frac{\sqrt{(1-f)\frac{s^2}{n}}}{\bar{y}} = \frac{\sqrt{\frac{s^2}{\bar{y}^2}}}{\sqrt{n}}\sqrt{(1-f)} = \frac{cv}{\sqrt{n}}\sqrt{(1-f)} \quad , \quad f = \frac{n}{N}$$

- **표본크기** : 유한모집단, 비복원추출을 가정
  - 모평균/ 모총합

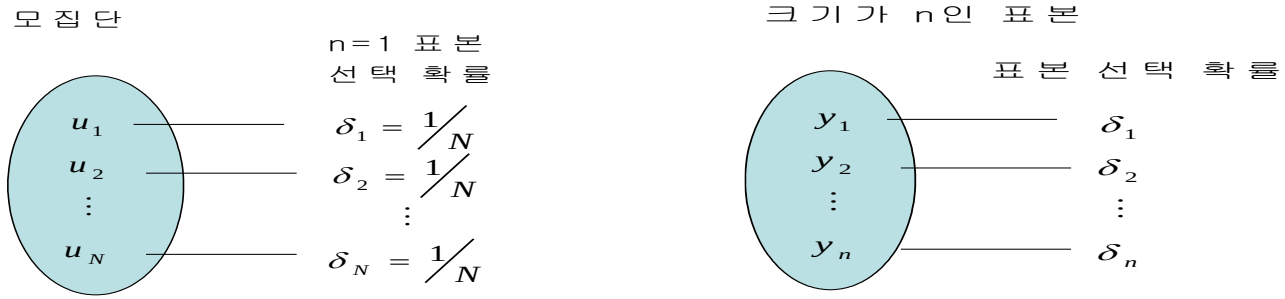
$$n = \begin{cases} \dfrac{N\sigma^2}{(N-1)D + \sigma^2} \\ \dfrac{Ns^2}{ND + s^2} \end{cases}, \quad D = \begin{cases} B^2/z^2 & , \text{모평균} \\ B^2/z^2N^2 & , \text{모총합} \end{cases}$$

  - 모비율/모총수

$$n = \begin{cases} \dfrac{NPQ}{(N-1)D + PQ} \\ \dfrac{N(0.25)}{ND + 0.25} \end{cases}, \quad D = \begin{cases} B^2/z^2 & , \text{모비율} \\ B^2/z^2N^2 & , \text{모총수} \end{cases}$$

[참고] 유한모집단에서의 표본추출과 추정

- 유한모집단에서의 표본추출 특성
  - 변동추출확률 : 표본추출확률이 모집단의 수와 추출 순서에 따라 표본단위의 추출확률이 변동
- WR case



  - 표본단위의 추출확률을 반영한 모총합 추정 $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\delta_i}$ with $\hat{\mathrm{V}}(\hat{\tau}) = \frac{s_{\hat{t}}^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{\delta_i} - \hat{\tau}\right)^2$
  - 추출확률의 결정
    - SRS : 모집단 크기만 고려
    - PPS : 관심변수 혹은 관심변수와 관계가 높은 보조 변수의 알려진 모집단 크기에 확률비례하도록 표본단위의 추출 확률을 결정(=the Probability Proportional to Size)

$$\delta_i = \begin{cases} \frac{1}{N}, & \text{SRS} \\ \frac{X_i}{\sum^N X_i}, & \text{PPS} \end{cases}$$

- WOR case
  - 특성
    - 표본단위의 추출확률은 모집단 크기 및 추출 순서에 따라 표본단위의 추출확률이 다름
    - N이 충분히 크고, n이 매우 작으면 WOR의 추출확률은 차이가 작으므로 WR의 추출확률과 근사적으로 동일하다는 설명이 가능
  - 변동추출확률의 계산

    $\pi_i = \Pr\{\text{unit (i) is selected in the sample}\}$ = 단위(i)가 표본에 포함될 확률의 평균
    - 표본단위의 추출확률대신 표본단위가 표본으로 포함될 확률(표본포함확률)로 반영
    - 추출확률과의 관계

$$\delta_i = \frac{\pi_i}{n} \Leftrightarrow \pi_i = n \cdot \delta_i$$

  - 총합추정량(Horvitz-Thompson estimator : HT 추정량)

$$\hat{\tau}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i \qquad \text{where} \qquad w_i = \frac{1}{\pi_i} \left(= \frac{1}{f}, SRS \right)$$

- 예 : 병원을 추출하는 경우
  - 추출방안

    1) SRS : 병원규모를 무시하고 동일한 추출확률로 추출하는 방안

    - N=10개이므로 1~10까지의 난수 중 3개(1, 2, 10)를 확률적으로 추출하여 표본 추출

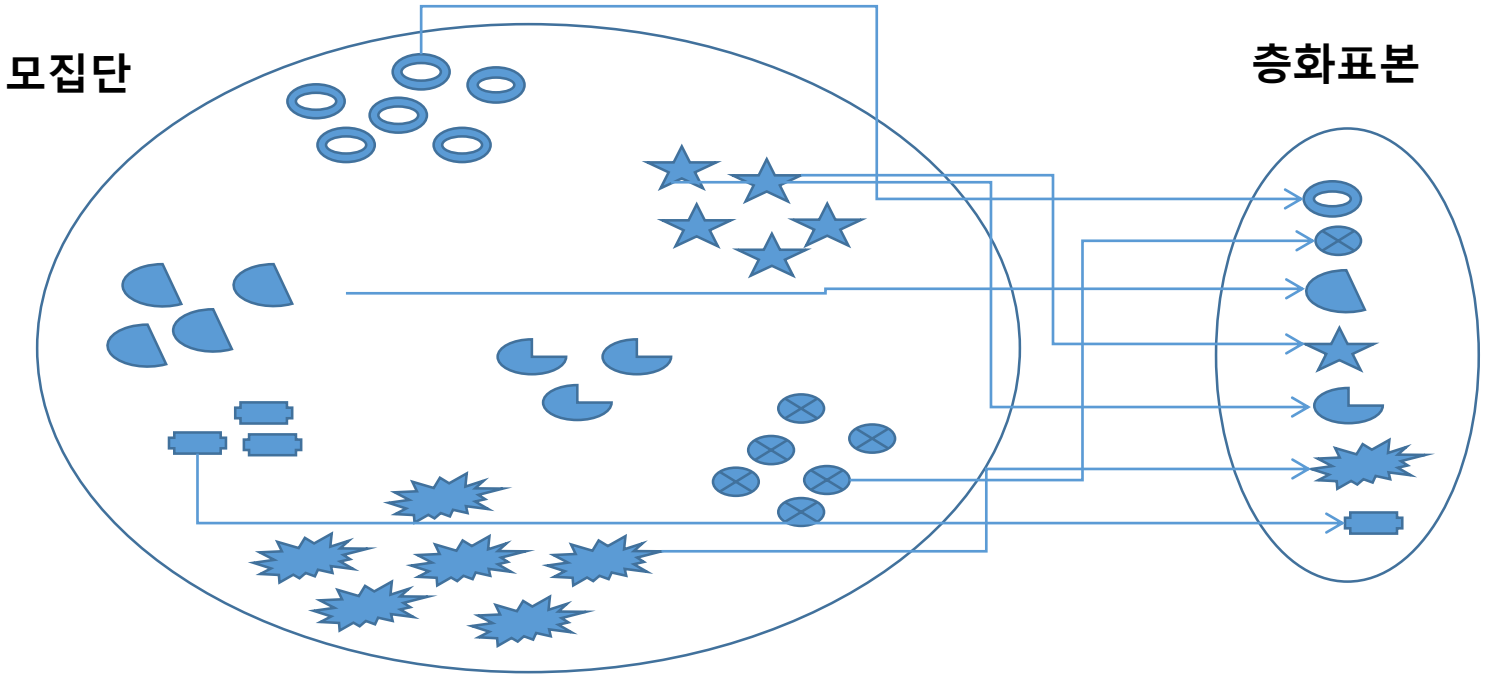
    2) SRS_PPS : 병원 의사수(혹은 환자수)를 반영해 병원마다 다른 추출확률로 추출하는 방안(의사수 비례 추출)

    - 총 의사수=328명이므로 1~328까지의 난수 중 3개(86, 201, 303)를 확률 추출하여 해당 병원을 표본으로 선정

id	의사수	누적합	매출액	추출률	표본	설계 가중치	
						PPS	SRS
10	91	91	72411	0.2774	O	1.2015	3.3333
1	128	219	53100	0.3902	O	0.8542	3.3333
8	39	258	23824	0.1189			
9	22	280	5800	0.0671			
6	6	286	4064	0.0183			
3	6	292	2797	0.0183			
2	13	305	2757	0.0396	O	8.4103	3.3333
4	4	309	2200	0.0122			
5	8	317	1950	0.0244			
7	11	328	1849	0.0335			
계	328	328	170752.1	1	3	10.4659	10

2) 층화추출법 (STS : stratified sampling)

- 정의 : 모집단 대상들이 이질적인 경우, 동질적인 그룹으로 층을 만든 후, 각 층마다 표본을 추출하는 방법



- 층화 기준
  - 층내 동질, 층간 이질적
  - 모집단 분할(partition)
- 층화 방법 : 군집분석 등 활용
- 층 수 : 이론적인 방법(Dalenius 정리, 제곱근 방법 등), 실용적인 방법

- 추출방법
  - 모집단 대상들을 동질적인 그룹으로 층화하고,
  - 각 층마다 모집단 대상(추출틀의 추출 단위)에게 일련번호를 부여하여
  - 각 층마다 확률장치(난수표, 난수 생성 프로그램)를 통해 n_h개의 번호를 무작위로 추출하여
  - 확률장치를 통해 추출된 번호와 일치하는 대상을 각 층의 표본으로 선정하여
  - 각 층마다 추출된 표본을 종합하여 n개의 층화 표본을 구성

- 특성
  - 모집단 대표성 증대 : 모집단이 이질적인 경우에 적절한 표본추출방법
    - 지역, 성별, 연령대 ; 병원규모, 진료과, 매출액 ; 사업체 규모, 산업분류 등
  - 층내 동질적, 층간 이질적인 층화 특징
    - 층내 단위의 추출 확률은 동일, 층간 단위의 추출 확률은 이질적 가능
    - 층 특성을 고려하여 층마다 서로 다른 추출법 적용 가능
  - 확률추출법의 기본 방법으로 다른 확률추출법의 기초 추출방법으로 활용
  - 장점
    - 모집단 대표성 증대로 표본오차 감소
    - 층별 추정 및 층간 비교 가능
    - 조사 관리의 편리성 증가
  - 단점 : 부모집단(층별 모집단) 크기 차이에 영향을 받음

- **추정식 : 모평균**
  - 모평균 추정 식: 표본평균

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h$$

  - 표본평균의 분산 추정량(WOR 가정)

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left(\frac{N_h - n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)$$

  - 추정의 정도
    . 추정오차한계(=신뢰구간의 폭)

$$\hat{B} = z\sqrt{\hat{V}(\bar{y}_{st})} \approx 2\sqrt{\frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \left(\frac{N_h - n_h}{N_h}\right)\left(\frac{s_h^2}{n_h}\right)}$$

    . 상대표준오차

$$\hat{R}SE = \frac{\sqrt{\hat{V}(\bar{y}_{st})}}{\bar{y}_{st}}$$

예) 2단계 층화추출법

$$\hat{\bar{Y}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij} \times y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}$$

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h - 1}\sum_{i=1}^{n_h}(e_{hi.} - \overline{e_{h..}})^2$$

여기서 $f_h = n_h / N_h$

$$e_{hi.} = [\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij} - \hat{\bar{Y}})]^2 / w_{..}$$

$$\overline{e_{h.}} = (\sum_{i=1}^{n_h} e_{hi.})/n_h$$

- 표본배분
  - 표본크기 n을 각 층으로 배분하는 기준 : $n_h = n \times W_h$
    - 각 층의 조사단위의 수 : 층의 부모집단 크기
    - 각 층의 변동 : 층내 분산
    - 각 층의 단위당 조사 비용
  - 표본배분 방법
    - 표본 배분은 기본적으로 층별 분석보다는 전체 추정을 목적으로 배분
    - 층별 분석이 목적인 조사에서는 최적배분, 네이만배분 및 비례배분은 적절하지 않음

최적배분	네이만배분	비례배분	균등배분	멱배분
$n_h = \frac{N_h s_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h s_h / \sqrt{c_h}} \cdot n$	$n_h = \frac{N_h s_h}{\sum_{h=1}^{L} N_h s_h} \cdot n$	$n_h = \frac{N_h}{\sum_{h=1}^{L} N_h} \cdot n = \frac{N_h}{N} \cdot n$	$n_h = \frac{1}{L} \cdot n$	$n_h = n \times \frac{(X_h)^p}{\sum (X_h)^p}$

    - 층간 분석 및 비교가 목적인 조사에서는 균등배분 및 멱배분(power allocation)을 널리 이용
      - 제곱근 비례배분

$$n_h = n \times \frac{\sqrt{N_h}}{\sum \sqrt{N_h}}$$

- 표본크기 결정

$$n = \frac{\sum_{h=1}^{L} N_h{}^2 \sigma_h{}^2 \Big/ w_h}{N^2 D + \sum_{h=1}^{L} N_h \sigma_h{}^2} \quad , \quad \mathrm{D} = \begin{cases} B^2 \Big/ z^2 & , \text{모평균} \\ B^2 \Big/ z^2 N^2 & , \text{모총합} \end{cases}$$

  - 여기서 w_h 층별 표본 배분 비율을 의미하며, $\sum_1^L W_h = 1$
  - 모비율은 $\sigma_h^2 = P_h Q_h$

[참고] 사후층화와 이중추출법

1. 사후층화

- 정의 : 층화 기준 변수에 대한 정보가 없거나 층화 관련 자료 획득이 불가능한 경우, SRS로 조사한 후에 추정단계에서 층화추출방법으로 추정하는 개념
  - SRS표본으로 조사한 자료가 모집단 구조와 상이할 때 사용
- 장점
  - 추정 정도가 향상
  - 모집단 및 부모집단 크기 N , N_h 이 알려져 있고, $n_h \geq 20$ 인 경우, 사후층화 추정량은 비례배분의 층화추출법의 결과와 비슷한 결과를 제공
- 추정량 : 모평균 추정 기준

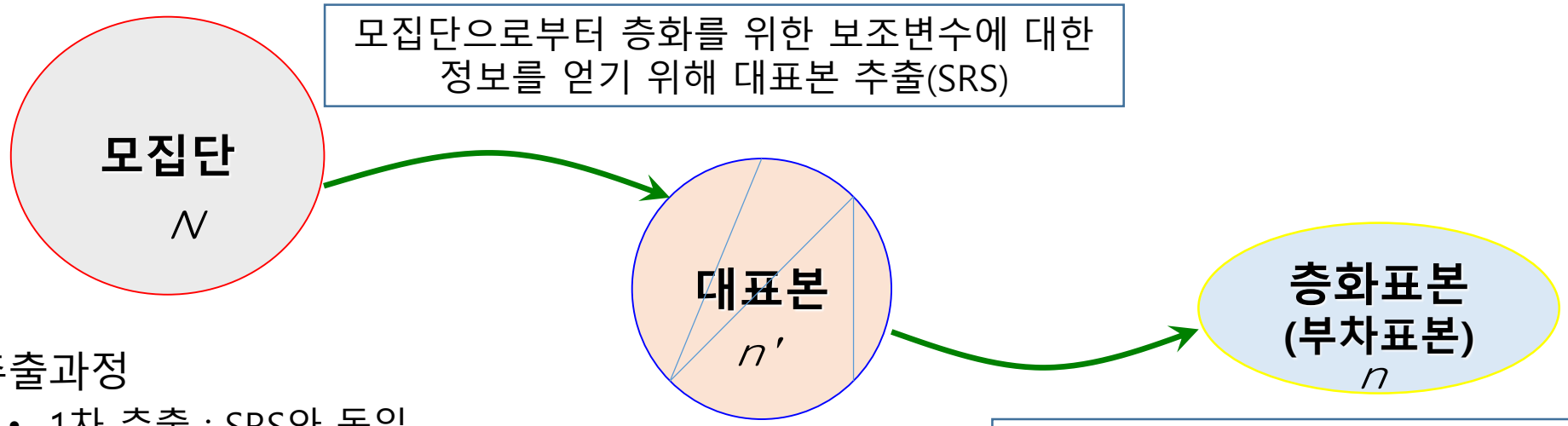
$$\bar{y}_{st\_ps} = \sum_{h=1}^{L} W_h \bar{y}_h \quad \text{, where} \quad W_h = \frac{N_h}{N}$$

$$\hat{V}(\bar{y}_{st})_{ps} = \frac{1}{n}\frac{N-n}{N}\sum W_h s_h^2 + \frac{1}{n^2}\sum (1-W_h)s_h^2$$

$$= \hat{V}(\bar{y}_{prop}) + \text{사후층화로 인한 분산}$$

2. 층화를 위한 이중추출법(double sampling)

- 기본 개념



대표본으로부터 획득한 사전 정보를 이용하여 조사 목적에 적합한 관심 사항을 기준으로 대표본에서 부차표본이 되는 층화표본을 추출

- 추출과정
  - 1차 추출 : SRS와 동일
  - 2차 추출 : 대표본을 모집단으로 하는 층화추출
- 추정량 : 모평균 추정

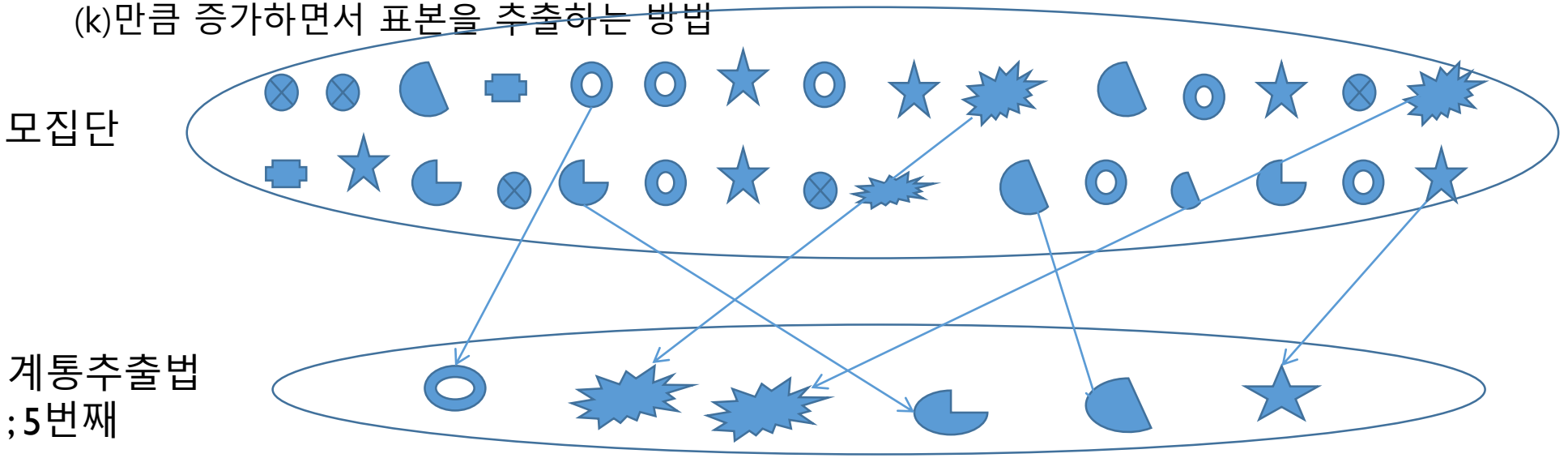
$$\bar{y}'_{st} = \sum_h w'_h \cdot \bar{y}_h \quad \text{where} \quad w'_h = \frac{n'_h}{n'}$$

$$V(\bar{y}'_{st}) = \frac{N - n'}{N} \cdot \frac{S^2}{n'} + \sum_h^L \frac{w'_h S_h^2}{n'} \left(\frac{n'_h}{n_h} - 1 \right)$$

$$= Var\left[E(\bar{y}'_{st} \mid sample_{\ 1})\right] + E\left[Var(\bar{y}'_{st} \mid sample_{\ 1})\right] = \text{1차추출분산} \quad + \text{2차추출분산}$$

3) 계통추출법 (SYS : systematic sampling)

- 정의 : 모집단 대상에게 일련번호를 부여하여 첫번째 표본을 무작위로 추출한 후, 두번째부터는 일정 간격 (k)만큼 증가하면서 표본을 추출하는 방법



- 추출 간격의 결정 : $k = \frac{N}{n}$
- 표본크기 $n = \frac{N}{k}$

- 추출방법
  - 모집단(추출틀) 대상에게 일련번호를 부여 : 가급적 관심변수와 선형적 관계가 되도록 부여
  - 추출 간격 k를 결정 : 모집단 크기 N을 모르면 근사적으로 결정
  - 1~k 사이에서 하나의 난수를 추출하여 첫 번째 표본으로 선정 : $1 \le r \le k$
  - 두 번째 이후부터는 추출 간격 k만큼씩 증가하면서 해당 번호의 대상을 표본으로 선정 $= r + (n-1)k$
  - 계통 표본 구성 : 일련번호 $\{r, r+k, r+2k, \ldots, r+(n-1)k\}$
- 특성
  - 표본추출이 편리하여 SRS 대신 사용 가능
  - 모집단 크기를 모를 때 적용 가능한 추출 방법 : 출구조사
  - 모집단 정렬순서가 관심변수와 무관하면 SRS와 추정 효율이 동일한 결과를 제공 : 추출오차 감소 효과
  - SRS보다 정도가 향상 : 순서모집단
  - 층화효과 :
    - 층마다 일정한 위치에 해당하는 하나의 표본을 추출한다는 의미 가능
  - 단점
    - 추정량의 분산 계산이 어려움 : 반복계통추출법
    - 추출틀의 형태에 의존 : 랜덤모집단, 순서모집단, 주기모집단

- **추정식 : 랜덤모집단을 가정하면, SRS와 동일**
  - 모평균 추정량인 표본평균의 분산 추정량은 편의추정량(biased estimator)
- **표본크기 : 랜덤모집단을 가정하면, SRS와 동일**

[참고] SRS와 SYS의 차이

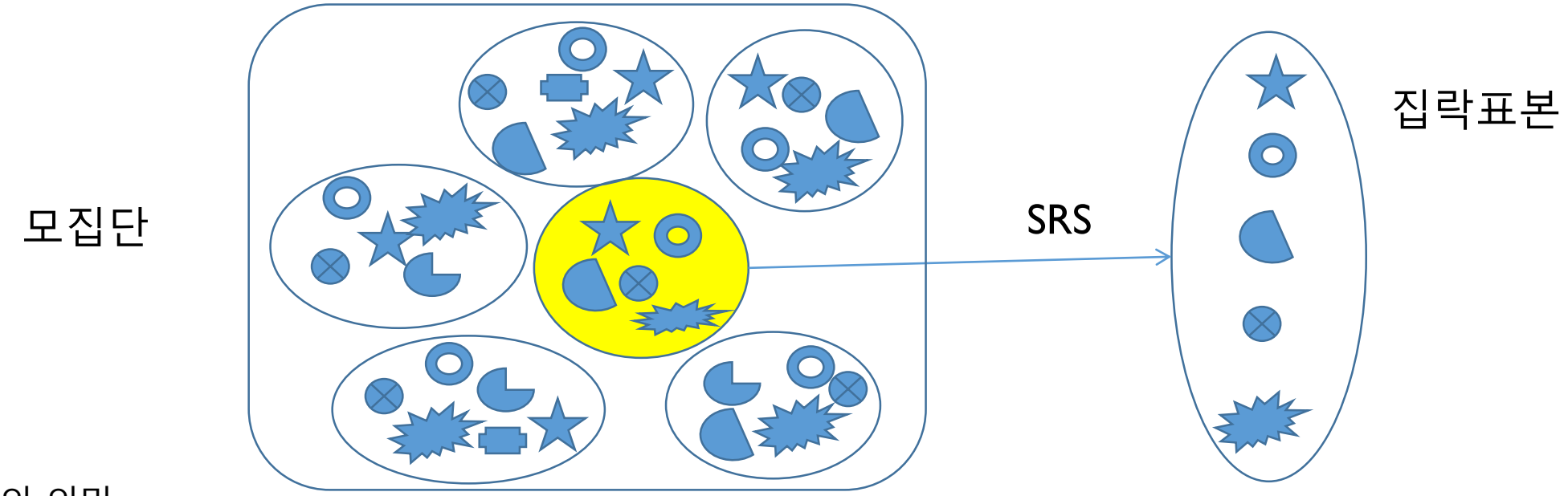
$$V(\bar{y}_{SRS\_WR}) = \frac{\sigma^2}{n}$$

$$V(\bar{y}_{SYS}) = \frac{\sigma^2}{n}\left[1 + (n-1)\rho\right] \quad \text{where} \quad \rho\text{는} \quad \text{급내 (계통표본내)상관계수 (= ICC)}$$

$$\text{여기서} \quad \rho = \begin{cases} \approx 0 & , \text{랜덤모집단} \\ < 0 & , \text{순서모집단} \\ > 0 & , \text{주기모집단} \end{cases}$$

4) 집락추출법 (CLS : Cluster sampling)

- 정의 : 서로 인접한 모집단 대상들로 구성된 집락(cluster)을 만들어, 집락을 표본으로 추출하여 집락내 대상을 표본으로 선정하는 방법(집락내 대상 전체 혹은 일부 추출이 가능)



- 집락의 의미
  - 지리적으로 인접한 모집단 대상들의 집단
  - 모집단과 유사한 아주 작은 집단
- 집락 구성의 원칙 :
  - 집락내 이질적, 집락간 동질적 : 집락내 상관정도가 영향
  - 집락의 크기 차이가 작게

- 추출 방법
  - 모집단(추출틀) 대상을 서로 인접한 대상들을 그룹화하여 집락을 구성 : 예) 인구주택총조사구
  - 집락에 대해 일련번호를 부여하고
  - 적절한 추출방법(SRS, SYS, PPS)으로 집락을 추출하여 표본 집락(1차 추출)을 추출
  - 표본 집락 내 대상 전체 혹은 일부를 추가 추출(2차 추출 : 2단계 집락 추출)하여 집락 표본을 구성

- 특성
  - 조사대상을 간접적으로 표본을 추출하는 방법
  - 추출틀 작성이 어려울 때 적절한 표본추출방법으로 널리 이용
  - 조사 편리성 증가
  - 조사 비용 감소
  - 단점
    - 표본오차가 증가 : 대부분 SRS보다 분산이 큼
    - 추정 결과는 집락내 상관정도의 영향을 크게 받음

- 집락표본의 설계 효과(design effect) : 집락표본의 설계 효과에 영향을 주는 요소
  - 집락 크기 : m (크기가 다르면 평균 크기를 이용)
  - 집락내 단위들의 동질성을 나타내는 급내 상관계수(ICC) : ρ
  - 집락 표본의 설계 효과

$$deff = D^2(\bar{y}_C) = 1 + (m - 1)\rho$$

- 급내상관계수(ICC : intra-class correlation)
  - 급내상관계수는 관심변수와 자연스럽게 구성된 집락의 특성에 의해 결정
  - 일반적으로 집락내 단위들이 동질적인 경향을 보이면, 급내상관계수는 양의 값을 가짐
  - 급내상관계수에 영향을 주는 요소 : 관심 변수, 집락 유형, 집락 크기
  - 사회과학 모집단에서 집락의 급내상관계수는 양수로 알려져 있음

- **추정식 : 1단계 집락추출법 가정**
  - **조사단위의 모평균 추정**

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

$$\hat{V}(\bar{y}) = \left(\frac{N - n}{Nn\,\overline{M}^2} \right) \frac{\sum_{i=1}^{n} (y_i - \bar{y} m_i)^2}{n-1}$$

$$2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{\left(\frac{N - n}{Nn\,\overline{M}^2} \right) \frac{\sum_{i=1}^{n} (y_i - \bar{y} m_i)^2}{n-1}}$$

  - 모집단 대상의 크기 M를 모르는 경우, 표본집락의 평균 크기 $\overline{m}$ 를 사용
  - 집락추출에서 모평균 추정량의 분산 추정량은 편의 추정량
  - 일반적으로 $n \geq 20$ 이면 좋은 분산 추정량을 제공하고, 집락크기가 동일하면 편의는 제거

- **표본크기 : 모평균 추정의 경우**

$$n = \frac{N\sigma_c^{\,2}}{ND + \sigma_c^{\,2}} \quad , \quad D = \frac{B^2\overline{M}^2}{4}$$

[참고] 집락추출법의 응용

- **층화집락추출법**
  - 집락을 층화하여 각 층마다 집락을 추출하는 방법
- **PPS에 의한 집락추출법**
  - 집락의 크기가 서로 다른 경우, 집락크기에 확률비례하도록 집락을 추출하는 방법
- **2단계 집락추출법**
  - 1단계에서 집락을 1차 추출하고, 2단계에서 표본 집락내에서 조사대상을 2차 추출하는 방법

[참고] 2단계 집락추출법

- **2단계 집락 추출법**
  - 표본추출방법
    - 1단계에서 표본집락을 추출
    - 2단계에서는 표본 집락내에서 조사 단위를 확률추출하는 방법
  - (예) 학교에서 n개 학급(psu)을 집락으로 추출하고 표본학급내에서 m명씩 조사단위를 추출행
    - 예로 120을 추출한다고 다양한 (n, m)의 조합이 존재 : 120 = n X m
    - 학급의 인원이 모두 동일한 경우 EPSEM 보장

$$\Pr(nm) = \Pr(n)\Pr(m \mid n) = \frac{n}{N} \times \frac{m}{M_n} = \frac{nm}{NM_n}$$

$$\Pr(nm) = \Pr(n)\Pr(m \mid n) = \frac{n}{N} \times \frac{m}{M} = \frac{nm}{NM}$$

  - 최적 (n, m)의 결정이 중요한 문제
    - 집락 및 조사단위 추출 비용을 고려하여 결정
    - 비용함수의 최소화 고려 $C = nc_1 + nmc_2$
    - m(=표본집락내 평균 추출 크기)의 결정

$$m = \sqrt{\frac{\sigma_w^2}{\sigma_b^2}\frac{c_1}{c_2}}$$ 여기서 b = between, w = within cluster

    - n의 결정 : 분산 최소화 혹은 비용 최소화 기준으로 결정

$$V(\hat{\mu}) = \frac{1}{n}\left[\sigma_b^2 + \frac{\sigma_w^2}{m}\right]$$